Predicting the Severity of Car Accident

Kedhar Natekar
September 19th 2020

1. Introduction to Business problem:

In the modern era of cities getting developed at a rapid pace and human migration happening towards the cities we examine that the roads which was meant for smooth traffic movement are getting congested. This is resulting in more accidents as the years pass by. Some are non-severe and some gets more severe.
This data set is the collection of such data of severity of accidents , time , location , road conditions , weather conditions , Accident type , type of collisions , people involved , vehicle involved etc.

We should be able to:

a. Predict the severity of an accident given the nature of accident.

b. Understand what factors lead to an accident to get Severe.

This helps in faster rescue and improving the conditions such that number of severe accident gets reduced.

Government owned agencies like fire department, police, ambulance services etc will be interested in understanding the severity of an accident given the parameters of how the accident occurred.
This helps them in faster implementation of their services knowing the factors that caused the accident. Also these agencies can improve the services such that severity of the accidents can be minimized in future.

2. Description of Data :

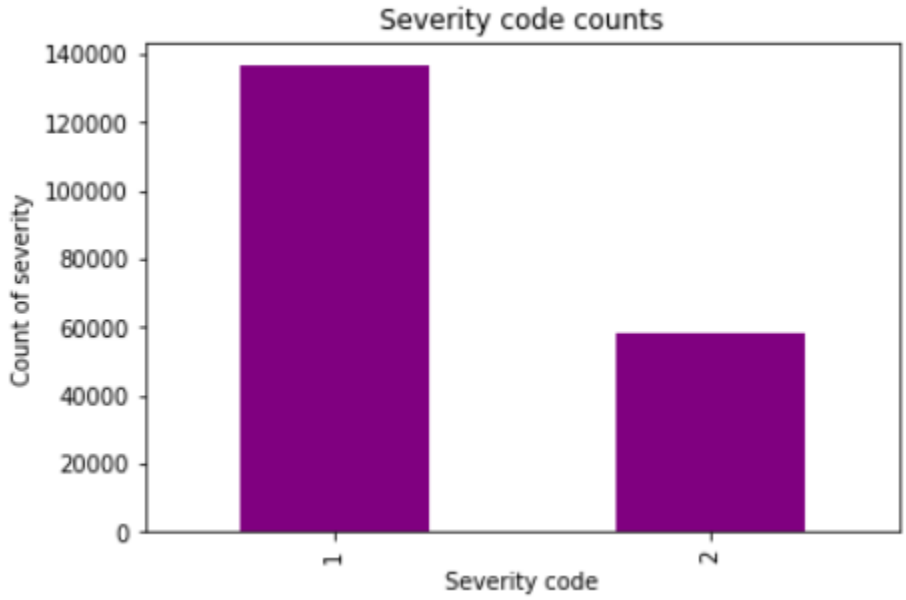
2. a) Data Sources :

Data which involves severity of accidents can we taken from official sources like open.canada.ca , data.gov.uk or kaggle sources. But the current dataset is sourced from the sample dataset given by Coursera IBM Data Science Capstone project.

2. b) Understanding the data:

The given data has 194673 reports of accidents.

The labelled column is SEVERITYCODE which has data in 1 and 2 which means 'prop damage' and 'injury'.



We have location co-ordinates along with location description which gives the exact street or location where accident has occurred.

Incident date and time is provided which is for past 20 years data when the accident occurred.

SDOT_COLCODE indicates the code in which it represent which vehicle involved in accident and the point of impact. Eg code of 14 indicates "MOTOR VEHICLE STRUCK MOTOR VEHICLE, REAR END".

Similarly we have ST_COLCODE which indicates movement of vehicles when the impact happened. Eg code of 14 indicates "From same direction - both going straight - one stopped - rear-end".

To understand the conditions when the accident occurred we have WEATHER, ROADCOND, LIGHTCOND which says the weather conditions, road conditions and light conditions.

Attributes like PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT and VEHCOUNT are numerical variables which gives the count of Persons, Pedestrians, Pedestrian cycles and Vehicles involved in the accident.

We also have some unique identifiers like ESRI unique identifier, Incident Keys.

3. Methodology used to solve the problem:

We have a huge data set and the target columns has either 1 or 2 are predictors which means it's a categorical variable and not continuous.

3. a) Model Selection:

Given the size of data and nature of prediction we used **Logistic Regression** to solve this approach.

SVM is ineffective with huge datasets and hence should be avoided.

KNN comes with huge computational cost and is ineffective with huge data set.

3. b) Statistical tests:

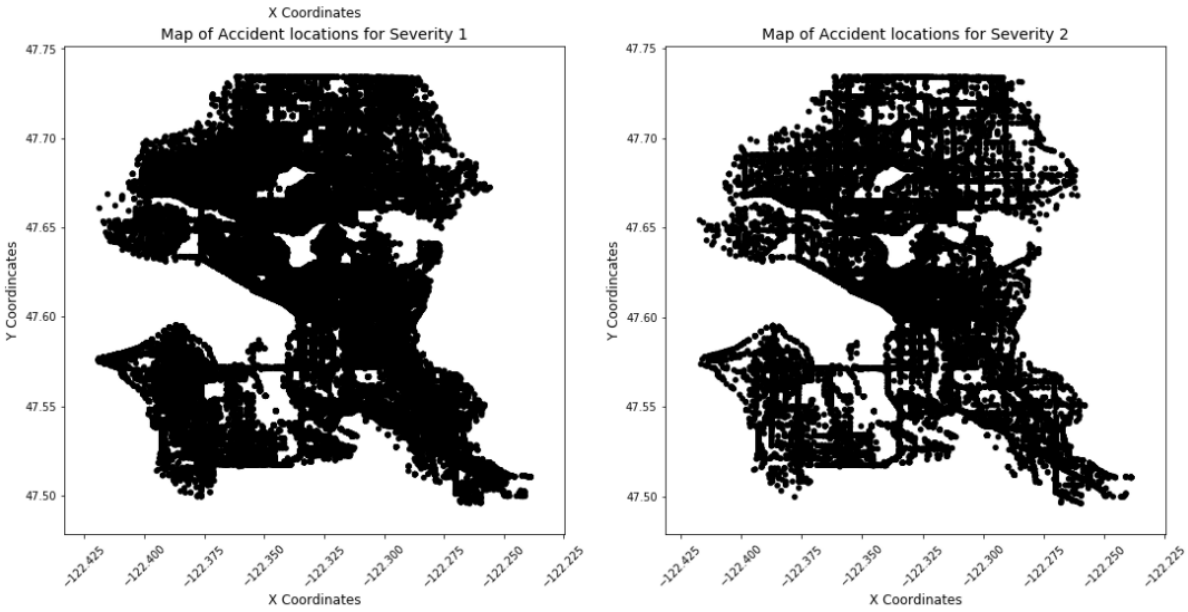
We have less continuous variables that can we used to predict categorical data. Hence Anova test for feature selection is used and the statistical significance is determined by p-value.
High F-value and p-value significantly minimum were considered

There are a lot of Categorical data and first we need to use dimensionality reduction to reduce the categories.
Then we used Chi2 test along with p-value to get the needful features out the given data.
High Chi2-value and p-value significantly minimum were considered.

3. c) Feature selections:

For location variables we need to understand is there any patterns in understanding which location has more severity impact. If yes we can divide the locations in categories of highly impactful and less impactful.

Below is the distribution for Severity codes 1 and 2 in different location.



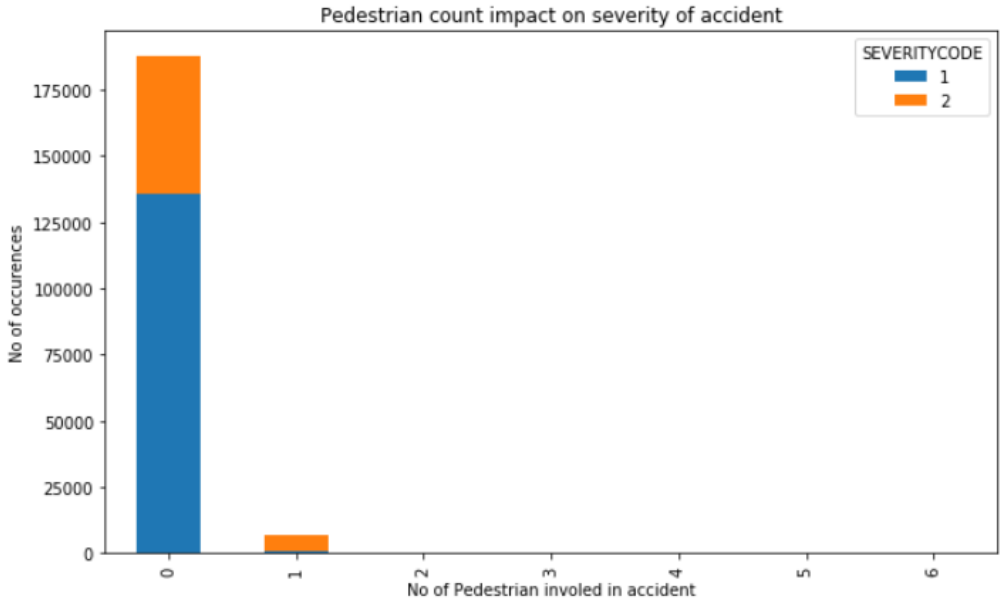
From above diagrams we see that Severity 1 and 2 are distributed evenly across and none of them is confined to any particular location. Hence we removed these X and Y coordinates along with Location attribute.

After performing the Anova and Chi-2 tests only features with best values were selected. There were 33 features including one-hot encoded variables and best 4 were selected as the one which gave better results.
Later only 2 features namely **PEDCYLCOUNT** (No of Bicycle's involved in accident) and **PEDCOUNT** (No of Pedestrians involved in accident) were selected as the final list of independent variables.

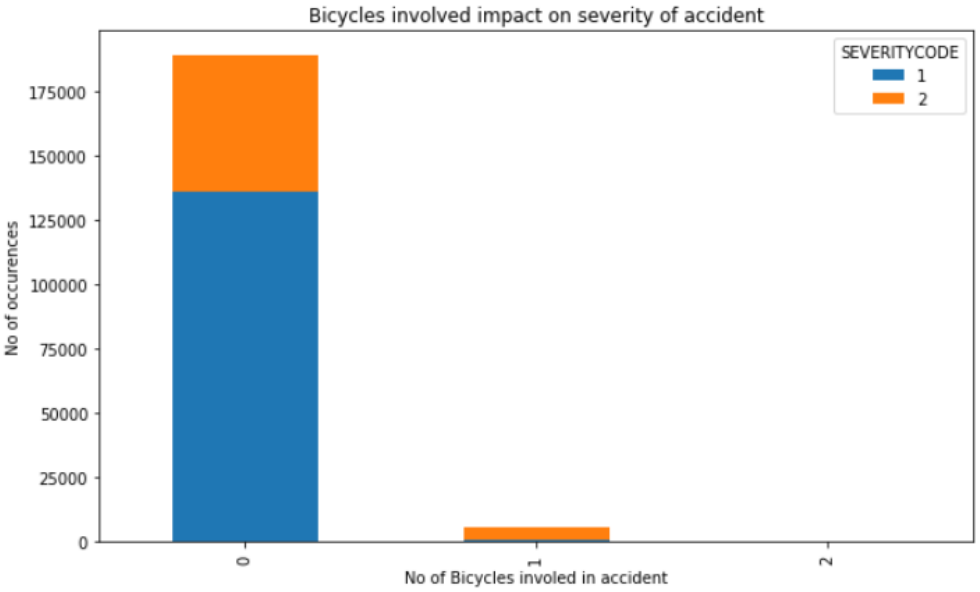
4. Results :

Knowing the no of Pedestrians involved in accident and no of bicycle's involved in accident gave more insight in predicting the Severity of an accident.

Below is how no of Pedestrians involved in accident affected its severity:



Below is how no of Bicycle's involved in accident affected its severity:



From above bar plots it's clear that if the no of pedestrians or the no of Bicycles that involve in an accident is more than 0 then the Severity of accident can be predicted as 2.

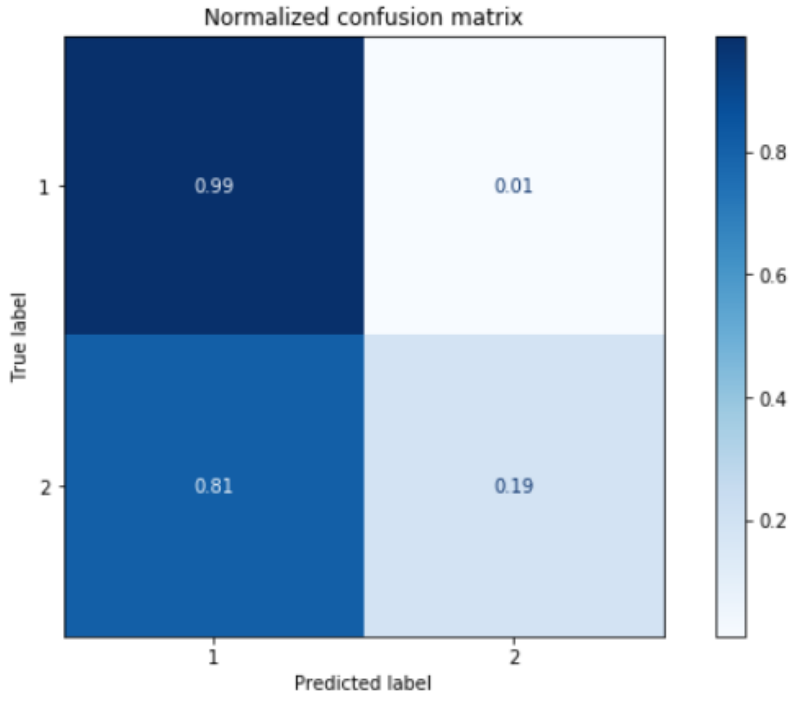
Accuracy, Log loss and Classification report of final model is as below:

```
Accuracy of the model is : 0.7385565775866761

Log loss of the model is : 0.5553687660937866

Classification Report given below :
```

	precision	recall	f1-score	support
1	0.74	0.99	0.85	27425
2	0.88	0.19	0.31	11510
accuracy			0.75	38935
macro avg	0.81	0.59	0.58	38935
weighted avg	0.79	0.75	0.69	38935

5. Understanding the results:

Confusion Matrix predicts that we are highly accurate is predicting if the accident severity is 1.



True positive for Predicting Severity of 1 is 99% affective which is a good sign.

But at the same time predicting if the severity is 2 is very less and 81% of such are predicted as 1.
In real world scenarios business will be more interested if the case is more severe as it's a matter of life or death.
The model should need more refinement such that it can predict 2 well with giving less importance to severity of 1.
Also data given has more references to severity 1 records than severity 2 and hence a poor prediction for severity 2. It will be good if we get more data on severity 2 so that we can predict it more accurately.

6. Conclusion:

We were able to drill down all the features in the given dataset and end-up the model with only 2 features with accuracy of 73.8%.
The model predicts 99% accurately if the severity is 1.
This model can be used by businesses to lookup for if the accident severity is 1.
There needs to be more refinement to predict the severity 2 and more such data can help in predicting more severe cases accurately.